LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-459331

# A primer for the statistical analysis of field test data obtained from screening instruments that detect contraband

D. R. Slaughter

October 13, 2010

## Disclaimer



This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

# Contents

# A primer for the statistical analysis of field test data obtained from screening instruments that detect contraband

## 1. Introduction

A variety of federal agencies have continuing development programs whose mission is to advance the technology for monitoring the flow of commerce to detect the presence of contraband. The items of interest can be undeclared goods on which taxes or tariffs are due, or illegal substances such as recreational drugs, or banned weapons or weapon parts such as nuclear material, or nuclear weapons being moved in violation of a non-proliferation treaty. As these technologies continue in development they are assessed periodically in experimental testing either in the laboratory or in the field. The testing programs sometimes may resemble those employed by the pharmaceutical industry and their goal is to determine whether these technologies are efficacious and whether there are unanticipated negative consequences. The basic questions are:

- Are they effective in detecting contraband?
- What are the negative consequences? What are the error rates?

Generally the first question is addressed by determination of the detection probability, $P_D$. The second is addressed by the false alarm probability, $P_{fA}$. These are the performance measures used to evaluate progress and to compare the performance of one technology or instrument against another. They also inform decisions regarding procurement of systems and their deployment to specific venues. The effectiveness of these instruments depends on the technology deployed, its calibration and stability, the CONOPS (conduct of operations) determining their usage, the nature of the contraband being detected, and the nature of the cargo overburden that may interfere with detection.

Testing results can be sorted into five categories, one of them ambiguous as follows:

1. **Detection**: Correct declaration that contraband is present when, in fact, there is contraband. The detection probability is $P_D$.
2. **Clear:** Correct declaration that contraband is not present when, in fact, there is no contraband. The clearing probability is $P_C$.
3. **False alarm:** Incorrect declaration that there is contraband present when, in fact, there is none. The false alarm probability is $P_{fA}$.
4. **Failure to detect**: Declaration that there is no contraband present when, in fact, there is contraband. The false negative probability is $P_{FN}$.
5. **No definite indication:** Some screenings will generate ambiguous results and fail to generate a definite indication of whether there is contraband present.

For the analysis below these cases will be separated from the main data set and not addressed in the analysis.

When screening systems are deployed to detect contraband their performance is characterized by $P_D$, $P_C$, $P_{fA}$, and $P_{FN}$ where

$$P_{FN} = \left[1 - P_D\right] \tag{1.1-1}$$

and

$$P_C = \left[1 - P_{fA}\right] \tag{1.1-2}$$

Unfortunately, the effectiveness of these systems in operation can be predicted only approximately since experimental testing produces only a somewhat ambiguous estimate of the true values for $P_D$ and $P_{fA}$. Their true values can be determined only in an infinite number of measurements. In practice an experimental campaign only approximates the infinite set and the result is an estimated value whose uncertainty is determined by the number of successes and sample size and the degree to which the experimental conditions accurately portray the real world implementation whose variations occur in materials, geometry, interferences, and deviations from prescribed procedures. This primer will deal only with the uncertainties due to stochastic variations in an otherwise perfect experimental simulation. The goal of the analysis is to determine the most probable value of P and estimate its uncertainty or the confidence that it falls within an acceptable range.

A subsidiary goal is to develop a good experimental plan that provides sufficiently robust statistical confidence so that, when the data analysis is completed, the performance uncertainties are acceptable and/or that the performance falls within the acceptable range at an acceptable confidence. It is important to estimate the minimum sample size that is likely to be required for robust uncertainty estimates so that the ultimate analysis produces compelling conclusions. The experimental results cannot be predicted in detail and so the minimum sample size cannot be predicted exactly. However, it is possible to determine the minimum size of a sample that contains no errors, or a sample with only one error, or only two, and so on. Those sample sizes are determined so that they provide the stated confidence level required in the analysis. So the goal is to define the number of errors at which testing will be stopped and to define the needed confidence, and then determine the sample size needed for that plan to be successful.

## 2 Two approaches to presentation of performance results

There are at least two approaches for presentation of the results of an experimental measurement of performance:

- A **point estimate** in which the most probable value of P is determined along with its uncertainty determined by the measured standard deviation, $\sigma_P$.

- A **one-sided confidence interval** or "**threshold estimate**" which establishes only the probability that the true value of P falls within an acceptable range. In the case of detection it establishes the probability that $P_D$ is in the acceptable range, i.e. $P_D(th) \le P_D \le 1$. Or, in the case of false alarms, the probability that the true value of $P_{fA}$ falls below a maximum allowed value, i.e. $0 \le P_{fA} \le P_{fA}(th)$.

The point estimate (developed in Section 5) is usually much more precise but, because of that, requires much more robust statistics and is more challenging experimentally. It usually requires large sample sizes to provide the desired result. On the other hand the one-sided confidence interval or "threshold" method (developed in Section 3) is somewhat ambiguous and forgiving, allowing the use of smaller data sets where only a few interesting events are observed. The latter is in common usage. For example the System Requirement for the Advanced Spectroscopic Portal (ASP) calls for detection performance of $P_D \ge 0.8$ at 95% confidence[1], meaning that there must be 95% probability that the true value of $P_D$ falls within the allowed range $0.8 \le P_D \le 1.0$. Similarly, the ASP specification sets an upper bound on false alarms on the neutron channel to $P_{fA} \le 0.001$, though no confidence is stated[2].

The following sections address both of these presentation formats for experimental data. And, more importantly, they establish the minimum sample sizes needed to confirm a given expectation of performance or requirement. The principal question is how many independent trials are likely to be required to establish that the true value of P falls within an allowed range at the stated confidence (probability that the true value is within that range) or to establish the most probable value of P with a satisfactorily small uncertainty.

There is ambiguity in forecasting the minimum sample size since the actual occurrence of errors (false alarms, fA, or failures to detect, FN) is not known beforehand. So the sample size estimates are based on hypothetical forecasts of the occurrence of errors in the data set. Sometimes the experimental protocol stipulates that testing is stopped when a predetermined number of errors have occurred. For that case the analysis presented

---

[1] Mike Kennedy, et. al., "Performance Specification for Advanced Spectroscopic Portal (ASP) Variant C (cargo)", US Department of Homeland Security Domestic Nuclear Detection Office (DNDO), 600-ASP-000013v4.10, July 19, 2007, requirement ASP-2862, page 6.

[2] Ibid, requirement ASP-1167, page 45.

below facilitates prediction of the number of test samples required based on the expected system performance.

# 3 The threshold estimate

## *3.1 The detection problem*

The true detection probability for a system in its normal operating regime is $P_D$. That can be measured exactly in a test with an infinite number of trials and is the ratio of successes to the total number of trials. But, actual field tests encompass only a finite set of trials and so the data provides only an estimate of $P_D$. How precise that estimate is depends on how closely the sample size approaches the infinite set of measurements. In real measurements that data, i.e. the number of trials and the number of successful outcomes, provides a most probable value of $P_D$ and a distribution function that describes the range of possible values and their relative likelihood. The methodology for this estimate may be found in many textbooks on statistical analysis. The best source is Wikipedia[3] and there are several traditional textbooks such as Box[4], Hoel[5], Mathews & Walker[6], or Feller[7].

For a random and independent set of N measurements that generate K correct detections the distribution function describing $P_D$ is the binomial distribution as follows[*]:

$$f(P_D,N,K) = \frac{N!}{K!(N-K)!} P_D{}^K \left(1 - P_D\right)^{N-K} \qquad (3.1\text{-}1)$$

The distribution of P is determined entirely by N and K. It's most probable value is always $\langle P_D \rangle = K/N$. The binomial distribution is shown in the figure below for several sample sizes, N, with nominally the same ratio $\langle P_D \rangle = 0.8$.

It is especially important to remember that this distribution describes the possible values of the true $P_D$ given a data set whose trials were genuinely random. There must be no correlations among the samples. Their sequence must be random to detect possible systematic errors and correlations among results due to common parameters such as utilizing the same test object repetitively in one cargo overburden or with the same interfering shielding or radioactive materials might introduce correlations that would violate the requirement for randomized trials. Certainly the data set must represent the

---

[3] http://en.wikipedia.org/wiki/Binomial_distribution

[4] Box, Hunter and Hunter, Statistics for Experimenters, Wiley (1978), page 130.

[5] Paul G. Hoel, Introduction to Mathematical Statistics, John Wiley & Sons (1971), page 59.

[6] Jon Mathews & R. L. Walker, Mathematical Methods of Physics, W. A. Benjamin (1964), page 354.

[7] William Feller, An Introduction to Probability Theory and Its Applications, Vol. 1, John Wiley & Sons (1968), pp 148 & 172.

[*] This is the standard form of the binomial distribution but is not normalized, i.e. its integral is 1/(N+1) so that in practical usage the numerator N! must be replaced by (N+1)! for proper normalization.

full diversity of the real world distribution of contraband items and their configuration in a shipment.

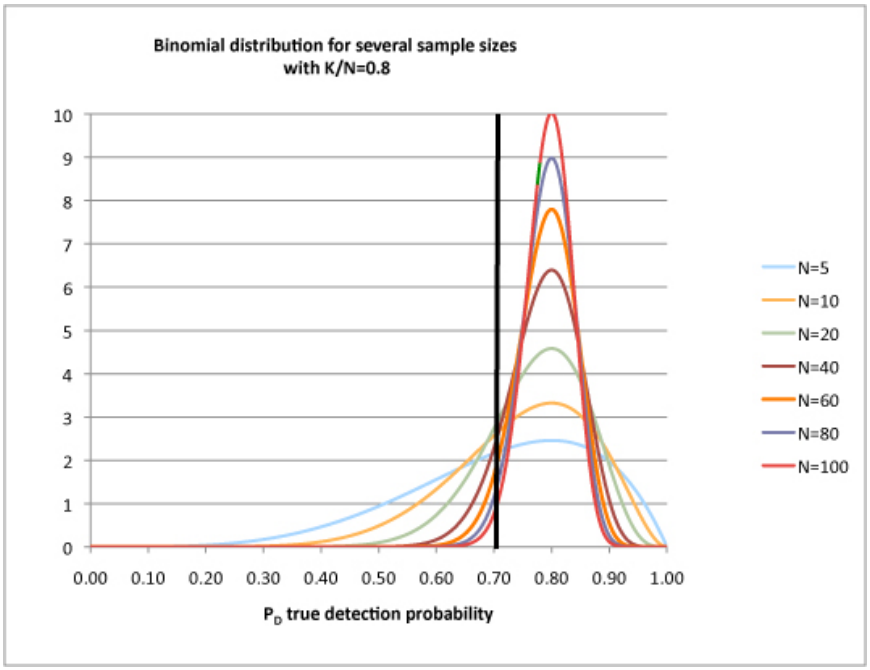


*Figure 3.1-1 Binomial distributions of various sample sizes with K/N=0.8.*

How narrow or broad this distribution is suggests the uncertainties in our estimate of $P_D$ and depends strongly on the sample size N. Clearly the breadth of the distribution and the magnitude of the uncertainties are greatest at small N such as N=5 and the distribution is seen to be asymmetrical about the most probable value for small N. When N becomes large, i.e. $N \geq 50$ or so the binomial distribution begins to resemble the Gaussian and that will later become the basis for a simple approximation.

Fortunately Eq. (3.1-1) can be easily evaluated in an Excel spreadsheet using the BINOMIAL function. Specifically*:

$$f(P_D,N,K)= (N+1)*\text{BINOMDIST}(K,N,P_D,\text{FALSE}) \qquad (3.1\text{-}2)$$

Evaluation of technology performance usually begins with estimation of the most probable value of $P_D$. That allows comparison of different technologies in the testing performance and a measure of their efficacy. However, all data sets are not the same. Some are a good deal more robust than others. It is important to assess the robustness of the estimated $P_D$ by analysis of the uncertainty in $P_D$. That uncertainty can be expressed as a standard deviation, as described in Section 5, but this is somewhat misleading when the distribution is highly asymmetrical as is the case in very small sample sizes. Alternatively, it is often more useful to establish a confidence interval within which the true vale of $P_D$ is likely to occur with high probability. The uncertainty is obtained from

* The factor (N+1) is required for normalization of the distribution function.

the one-sided confidence interval or "threshold" approach where, rather than determining the most probable $P_D$ alone, the method determines the probability that $P_D$ exceeds a threshold value. What is the probability that the true value of $P_D$ exceeds a minimum required value, i.e. is to the right of the vertical bar in Fig. 3.1-1? This probability is called the "confidence", i.e. the probability that the true $P_D$ falls within the acceptable range, $P_D(th) \le P_D \le 1.0$. The confidence integral is given below.

$$C(P_D,N,K) = \int_{P_D(th)}^{1} f(P,N,K)dP \qquad (3.1\text{-}3)$$

In the example above the most probable value was $<P_D>=0.8$ and a threshold value was set to $P_D(th)=0.7$. For that threshold value the confidence integral, Eq. 3.1-3, takes on the following values.

| N | K | Errors, m=N-K | Confidence, C (%) |
|---|---|---|---|
| 5 | 4 | 1 | 57 |
| 10 | 8 | 2 | 67 |
| 20 | 16 | 4 | 79 |
| 40 | 32 | 8 | 89 |
| 60 | 48 | 12 | 94 |
| 80 | 64 | 16 | 97 |
| 100 | 80 | 20 | 98 |

As the figure illustrates and as the table shows the confidence for the interval increases dramatically with sample size as the distribution grows narrower. The distribution becomes more compact about its most probable value and becomes more symmetrical. At N=5 there's a 43% probability that $P_D$ falls outside the acceptable range while at N=100 there's only 2% probability that $P_D$ is outside the acceptable range.

Fortunately, the confidence integral is available in an Excel spreadsheet using the same BINOMIAL function. The last argument is switched to accomplish this as follows*:

$$C(P_D(th),N,K) = \text{BINOMDIST}(K,N+1,P_D(th),\text{TRUE}) \qquad (3.1\text{-}4)$$

The confidence, i.e. the probability that the true $P_D$ falls within the defined interval, depends strongly on the threshold value as shown in the figure below

* Just as before the argument N must be replaced by N+1 to provide proper normalization.

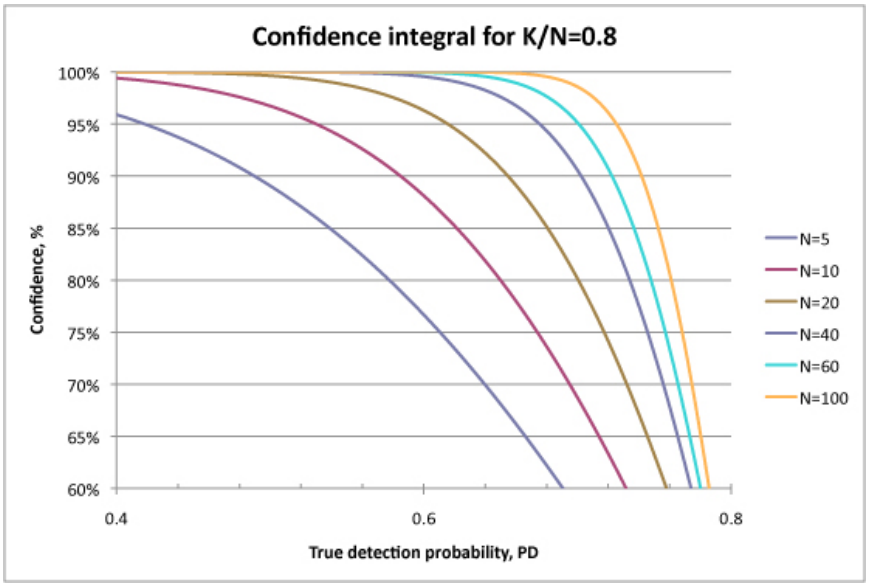


*Figure 3.1-2 Confidence integral for K/N=0.8.*

It is clear that estimation of the minimum sample size, $N_{min}$, depends on the expected performance, $\langle P_D \rangle = K/N$ (how many errors are likely to occur in a sample set), and the desired confidence at the threshold value $P_D(th)$. There is no analytic solution above and, instead, it is necessary to solve Eq. (3.1-3) above iteratively to determine N given C. That is, the threshold $P_D(th)$ and the confidence C are set to predetermined values, then N and K are varied to satisfy the integral equation. An example is shown below where the minimum samples sizes have been estimated for several values of $P_D$ and at two confidence levels.

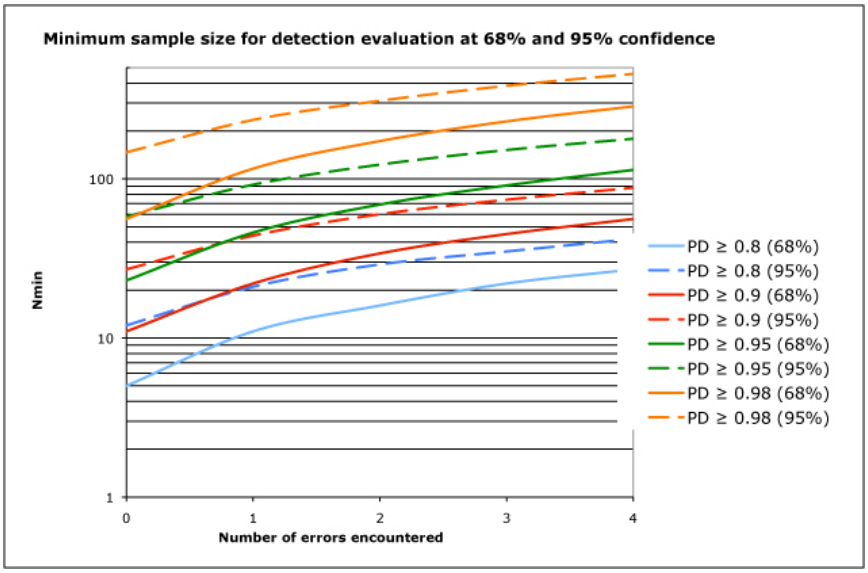


*Figure 3.1-3 Minimum sample size vs. the number of errors encountered. Two cases: confidence = 68% and confidence = 95%.*

The figure above shows that the demand for increased sample size grows dramatically with increases in the needed confidence. Of course additional errors generate a need for larger samples to confirm a given detection probability at the same confidence.

In general there's no quantitative knowledge of $P_D$ until a few errors have been encountered in the measured data set. It's not possible to predict when the errors will occur or how many trials will occur before the first error is encountered, or the second, or the third, and so on. Instead, the problem is posed in terms of the minimum sample size required if no errors are encountered, the minimum size if only one is encountered, or if only two or only three or only four errors are encountered. So, the minimum requirement shown in Fig 3.1-3 is used by selecting a "few" errors, hypothetically establishing a given value of m, where $1 \leq m \leq 4$, and determining the needed sample size from the figure for that hypothetical result. Then if the data set contains a few errors and the minimum sample size has been reached in the data for that number of errors, then the resulting N and K will be sufficient to confirm that the performance is within the acceptable range at the defined confidence, or that the uncertainties are adequately small.

## *3.2 Clearing probability*

The "clearing probability", $P_C$, is completely analogous to detection where a cargo with no contraband present is correctly cleared. This is a positive outcome. It occurs with high probability and is analyzed in the same way as detection by replacing $P_D$ with $P_C$ in Eq. (3.1-3). Thus, the clearing probability is given below.

$$f(P_C,N,K) = \frac{N!}{K!(N-K)!} P_C{}^K \left(1 - P_C\right)^{N-K} \tag{3.2-1}$$

and its confidence integral is given by:

$$C(P_C,N,K) = \int_{P_C(th)}^{1} f(P,N,K)dP \tag{3.2-2}$$

In practice the usual approach is to work the inverse problem, i.e. the false alarm probability. False alarms are the complement to correct clearing.

$$P_C = \left(1 - P_{fA}\right) \tag{3.2-3}$$

The relationship noted above is utilized in the analysis of false positives or false alarms below.

## 3.3 False alarm analysis

Analysis of false alarm probability, $P_{fA}$, is identical to the approach above for detection probability, $P_D$. In this case an upper bound, $P_{fA}(th)$, is set to a predetermined value (shown by the bar in Fig 3.3–1, that sets a limiting value on $P_{fA}$ and the range of interest is over the small values of $P_{fA}$ ranging $0 \le P_{fA} \le P_{fA}(th)$.

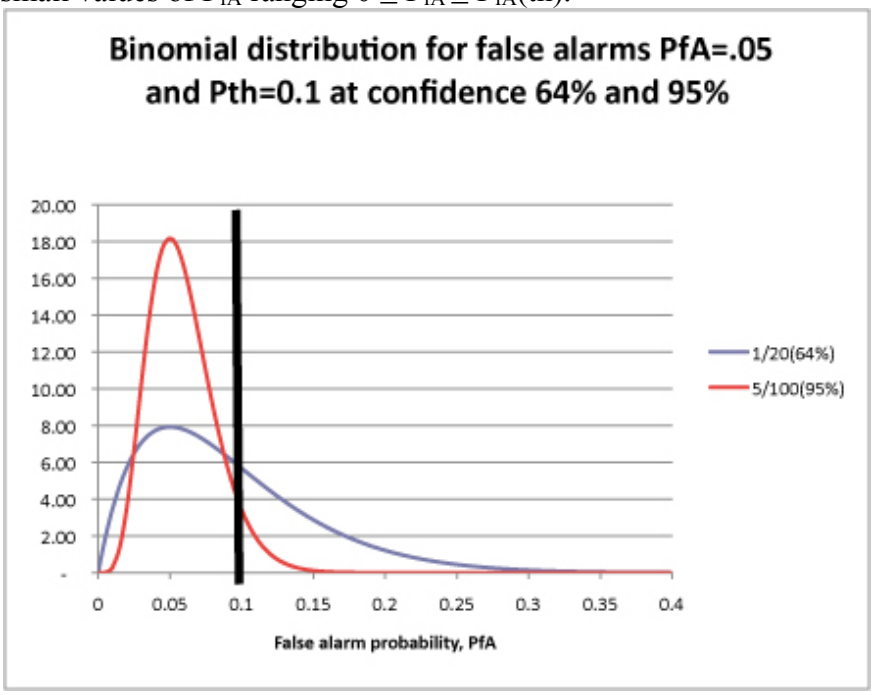


*Figure 3.3-1 Binomial distribution for $P_{fA}$=0.05. Curves are for K/N=1/20 and K/N=5/100. The former corresponds to 64% confidence that $P_{fA} \le 0.1$ and the second to 95% confidence that $P_{fA} \le 0.1$.*

The binomial distribution peaks (most probable value) at 0.05 for 1 error in 20 trials and the distribution of likely $P_{fA}$ values extends above the threshold value $P_{fA}(th)$=0.1 with low probabilities. Here the confidence integral is re-cast into the following form:

$$C(P_{fA}(th),N,K) = \int_0^{P_{fA}(th)} f(P,N,K)dP \qquad (3.3\text{-}1)$$

In the figure example two data sets are considered both of which have a most probable $P_{fA}$=0.05. However, one set contains only 20 trials and the other 100 trials so that the larger data set has a much narrower distribution and more of it falls in the acceptable range $P_{fA}(th) \le 0.1$. Clearly the larger sample size leads to higher levels of confidence that the true value of P is within the acceptable range. For the N=20 set the confidence is only 64%, i.e. there is a 36% probability that P is not within the acceptable range below $P_{fA}(th) \le 0.1$. But, for the N=100 set the distribution is much narrower so that the confidence is 95% indicating only a 5% probability that the true $P_{fA}$ is outside the acceptable range.

The confidence integral for this case is shown below.

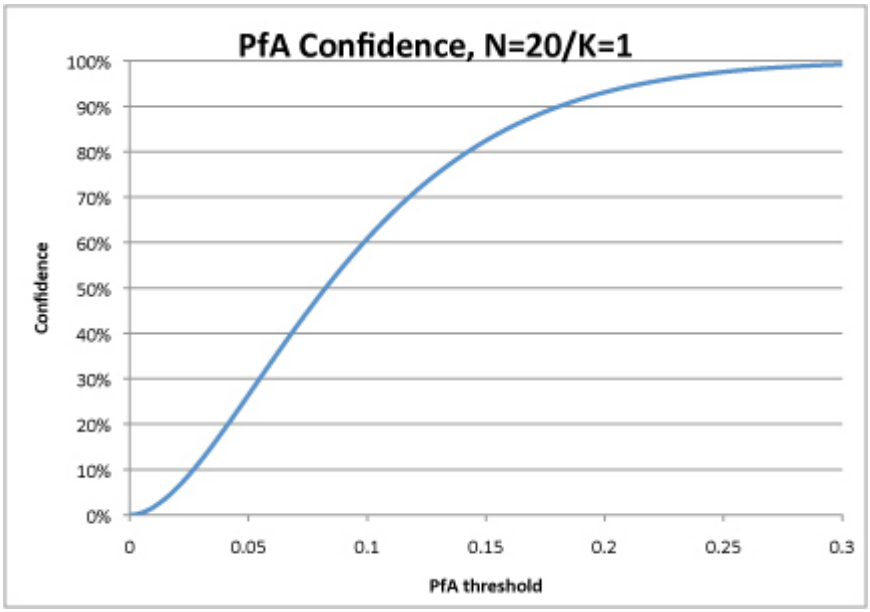


*Figure 3.3-2 Confidence integral for N=20 trials with K=1 false alarm.*

In general the expression is evaluated in a manner analogous to the detection problem and the result for the inverse is derived from Eq. (3.3-1) above:

$$C(P_{fA}(th),N,K) = 1 - \int_{P_{fA}(th)}^{1} f(P,N,K)dP \qquad (3.3\text{-}2)$$

and the spreadsheet version is given below*:

$$C(P_{fA}(th),N,K) = [1 - \text{BINOMDIST}(K,N+1,PfA(th),TRUE)] \qquad (3.3\text{-}3)$$

This equation is solved iteratively to determine the minimum sample size exactly as in the detection problem. An example of the results is shown in the figure below.

* As before replacing N with N+1 provides needed normalization.

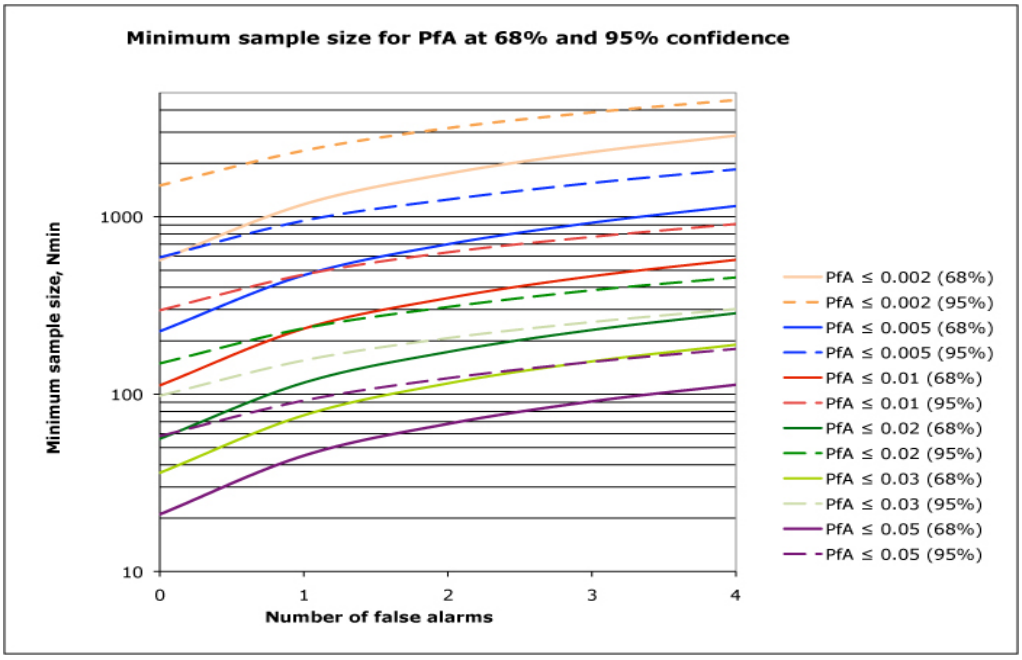


*Figure 3.3-3 Estimated sample size vs. number of false alarms for two levels of confidence and $P_{fA}$ thresholds 0.002 to 0.050.*

Here too it is seen clearly that the number of trials required increases with the desired confidence and with the number of observed alarms. A quantitative knowledge of $P_{fA}$ is obtained only after a few errors have been encountered in the data set. The figure above is used by choosing a "few" errors, i.e. $1 \leq m \leq 4$ and determining the minimum sample size from the figure based on the number of errors that could be in the data set when testing is stopped.

The procedure is exact but can be time consuming. An alternative that offers an analytic solution that is simple to calculate is desirable and an approximation provides this convenience is described below.

## *3.4 Failure to detect*

Failure to detect is completely analogous to the false alarm. It is an error condition, i.e. a negative outcome that occurs with low probability. The failure to detect is described by replacing $P_{fA}$ with $P_{FN}$, the probability of a failure to detect or "false negative".

# 4 The analytic (Gaussian) approximation

## 4.1 Rough estimation of the needed sample size

The methodology described above is correct in its description of the statistics but it is cumbersome to evaluate. An analytical formula would be more convenient. When sample sizes are large the Gaussian distribution is a reasonable approximation to the binomial and it facilitates an analytical formula. The Gaussian description of the detection probability is given by:

$$f(K,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{K-\mu}{\sigma}\right)^2} \qquad (4.1\text{-}1)$$

where for sample size N the distribution function has:

Mean: $\mu = PN$ (4.1-2)

Standard deviation: $\sigma = \sqrt{\mu} = \sqrt{PN}$ (4.1-3)

The Gaussian shape is very familiar where its smooth tails extend to infinity in both directions and the distribution is symmetric about the mean. The most probable value occurs at the mean. Comparison with Figures 3.1-1 and 3.3-1 show stark contrast between the symmetrical Gaussian shape and the highly asymmetrical shape of the Binomial distribution. However, these differences are reduced with increasing sample size as can be seen in Figure 3.1-1 for N = 100. For a more detailed discussion of the Gaussian approximation and its errors, and its range of applicability see Wikipedia[8].

## 4.2 Detection

The simplest formulation of the Gaussian approximation is to analyze the probability of a false negative result. That is the inverse of the detection probability, i.e.

$$P_{FN} = \left[1 - P_D\right] \qquad (4.2\text{-}1)$$

The false negative probability is then given, in this approximation, by:

$$P_{FN}(K,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{K-\mu}{\sigma}\right)^2} \qquad (4.2\text{-}2)$$

where K is the number of false negatives, $\mu$ is the mean number of false negatives and $\sigma$ is the standard deviation of the distribution of false negatives.

[8] http://en.wikipedia.org/wiki/Binomial_distribution

$$\mu = P_{FN}N \tag{4.2-3}$$

$$\sigma = \sqrt{\mu} = \sqrt{P_{FN}N} \tag{4.2-4}$$

The probability that the number of false negatives is less than the threshold $P_{FN}(th)$

$$C(K,\mu,\sigma) = \int_{-\infty}^{P_{FN}(th)} P_{FN}(K,\mu_{FN},\sigma_{FN})dK = \frac{1}{2} + \int_{0}^{P_{FN}(th)} P_{FN}(K,\mu_{FN},\sigma_{FN})dK \tag{4.2-5}$$

The confidence integral is evaluated by defining a new variable t as follows:

$$t = \frac{(K - \mu_{FN})}{\sigma_{FN}\sqrt{2}} \tag{4.2-6}$$

and

$$t_{th} = \frac{(K_{th} - \mu_{FN})}{\sigma_{FN}\sqrt{2}} \tag{4.2-7}$$

The confidence is established by integration over the Gaussian probability distribution. Then taking the definition of the Error Function[9], i.e.

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} dt \tag{4.2-8}$$

The result is:

$$C(K_{th},\mu_{FN},\sigma_{FN}) = \frac{1}{2}\left[1 + erf(t_{th})\right] \tag{4.2-9}$$

The error function is tabulated in standard texts such as AMS-55[10]. For convenience a new parameter Z is defined:

$$Z_{th} = t_{th}\sqrt{2} = \frac{K_{th} - \mu_{FN}}{\sigma_{FN}} \tag{4.2-10}$$

Then several values for C in Eq. (4.2-9) are chosen and $Z_{th}$ is determined that satisfies the equation. The results are shown in the table below.

---

9 P. M. Morse, et. al.,"Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables", National Bureau of Standards, Applied Mathematics Series #55, (1964), page 297.

10 Ibid, page 310.

*Table 4.2-1 $Z_{th}$ for several levels of confidence*

| Confidence, C (%) | erf ($t_{th}$) | $t_{th}$ | $Z_{th}$ |
|---|---|---|---|
| 68 | 0.36 | 0.33 | 0.47 |
| 80 | 0.60 | 0.60 | 0.85 |
| 90 | 0.80 | 0.91 | 1.29 |
| 95 | 0.90 | 1.16 | 1.64 |
| 98 | 0.96 | 1.45 | 2.05 |
| 99 | 0.98 | 1.645 | 2.33 |

In a given measurement data set there will be "m" false negative errors and, in the Gaussian approximation, that constitutes both the mean and most probable value of the distribution function, i.e.

$$\mu = m \tag{4.2-11}$$

In addition, the Gaussian approximation assumes the standard deviation:

$$\sigma = \sqrt{\mu} = \sqrt{m} \tag{4.2-12}$$

Combining these results from Eqs. 4.2-10, 4.2-11, and 4.2-12 and noting that $K_{th}=P_{FN}(th)xN$ Eq. 4.2-10 is solved for N.

$$N = \frac{m + \sqrt{m}Z_{th}}{P_{FN}(th)} \tag{4.2-13}$$

But $P_{FN}$ is the complement of $P_D$ (from Eq. 4.2-1) so this result is written in the more appropriate form below.

$$N_D = \frac{\left(m + Z_{th}\sqrt{m}\right)}{\left(1 - P_D(th)\right)} \tag{4.2-14}$$

This is the desired result when analyzing **detection data**, i.e. when real contraband is present. For that analysis the result gives the estimated minimum sample size in the Gaussian approximation. It may be compared to the exact result from the binomial distribution and that is shown below.

Utilization of this result relies on the following understandings:

- $P_D(th)$ is the lower bound of the desired confidence interval. It's the lowest acceptable value for $P_D$.

- m is the number of errors at which testing is stopped. Usually a few errors would be required to establish confidence in the knowledge of $P_D$. However, the confidence interval is a relatively broad range so that a large number of errors, m, is not required. Usual judgments would place appropriate values for m in the range $1 \leq m \leq 4$.

- $N_D$ is the minimum sample size that is expected to be required in order to confirm that the true value of $P_D$ is within the desired range at the stated confidence (reflected in $Z_{th}$).

The approximate analytical result above is compared to the exact result in the figure below.

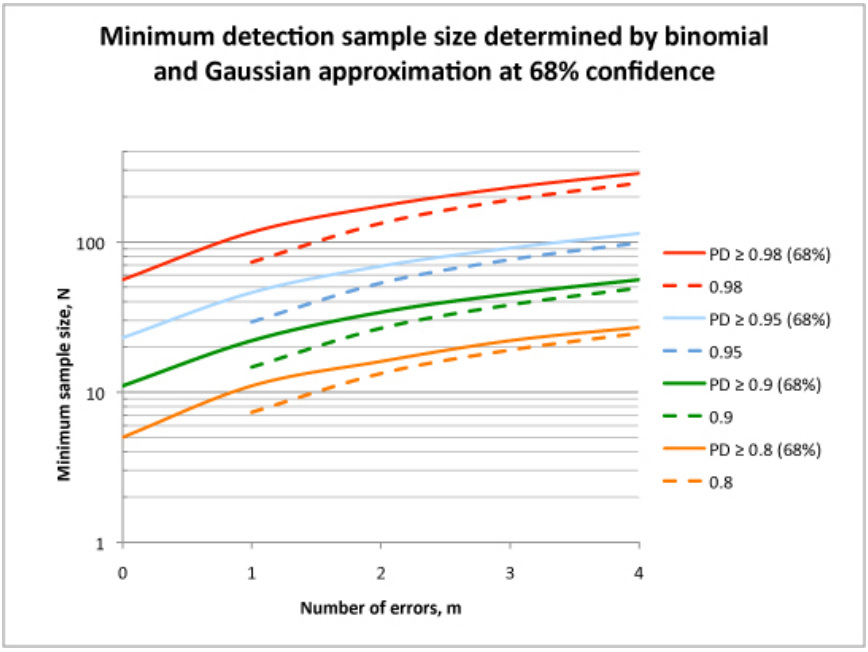


*Figure 4.2-1 Comparison of Gaussian formula estimate of sample size to exact value from binomial distribution. Solid lines are the binomial determination of minimum sample size while the dashed curves show the Gaussian estimate.*

Examination of the figure shows that the simple result derived for the Gaussian approximation at 68% confidence provides sample size estimates quite close to the correct values obtained from the binomial. The accuracy of the prediction improves with increasing m. This result appears to be applicable over a substantial range of detection thresholds at 68% confidence that spans nearly two decades in sample size. The estimates are better when the threshold value is high as compared to the cases with low threshold values. Generally the analytical estimate in Eq. 4.2-14 underestimates the exact result from the binomial by only 12-35 % in the regime focused on 68% confidence. The greatest errors are, as expected, at the extreme where there is only 1-2 false negatives. This is considered a good estimate in light of the very poor representation of the distribution function by the Gaussian at small error rates.

A similar comparison can be made for the Gaussian estimates at 95% confidence. That result is shown below.

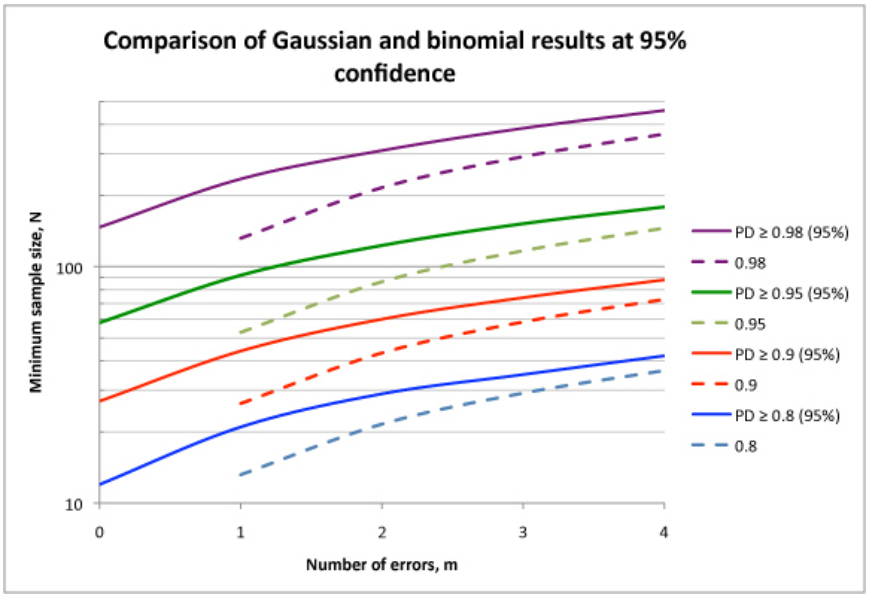


*Figure 4.2-2 Comparison of Gaussian approximation estimate of sample size to binomial at 95% confidence. Solid curves are the sample size determined from the binomial while the dashed lines show the Gaussian estimate.*

Examination of the figure shows the Gaussian estimate to be fairly precise over the range investigated when the confidence is 95%. Accuracy improves at larger m. The Gaussian approximation generates sample size estimates that are slightly low, possibly as much as 20-40 % low. This is considered a reasonably good approximation as the scale of the underestimate remains small over two decades in sample sizes.

## 4.3 False alarms

The Gaussian approximation in the case of false alarm is analogous to the approach for detection. With the focus on the error condition, i.e. the occurrence of false alarms. The confidence integral is obtained from Eq. (4.2-9) where the parameter t is defined with a mean equal to the number of false alarms occurring in a given data set and the standard deviation is the square root of the mean. The threshold in this case is:

$$K_{th} = P_{fA}(th)N \tag{4.3-1}$$

Then from Eq. (4.2-10)

$$Z_{th} = \frac{\left(P_{fA}N - m\right)}{\sqrt{m}} \tag{4.3-2}$$

From which the desired result is:

$$N_{fA} = \frac{m + Z_{th}\sqrt{m}}{P_{fA}(th)}$$

(4.3-3)

This provides the desired result when **analyzing false alarm data**, i.e. data obtained when no contraband is present. The result is compared in the two figures below with the exact result obtained from the binomial distribution.

This result is utilized with the following understandings:

- $P_{fA}(th)$ is the threshold value establishing the confidence interval. It is the largest acceptable value of $P_{fA}$.
- $P_{fA}$ has a high probability of falling within the acceptable range. That probability or confidence is reflected by $Z_{th}$.
- m is a judgment choice reflecting the number of errors likely to be encountered in the data set before the trials are stopped. Good judgment usually requires a few errors before the estimates are considered good so $1 \leq m \leq 4$.
- Then $N_{fA}$ is the minimum sample size that is expected to be required to confirm that $P_{fA}$ falls within the acceptable range at the confidence embedded in $Z_{th}$.
- If these ground rules are adhered to then the completed data set with $N_{fA}$ samples would subsequently be analyzed using the binomial distribution (3.1-3) and the sample size is expected to be sufficient to confirm $P_{fA} \leq P_{fA}(th)$ at the defined confidence.

The analytical result is compared to the binomial in the figures below. The extent to which the Gaussian approximation is adequate can be seen comparing the solid and dashed curves in the figures.

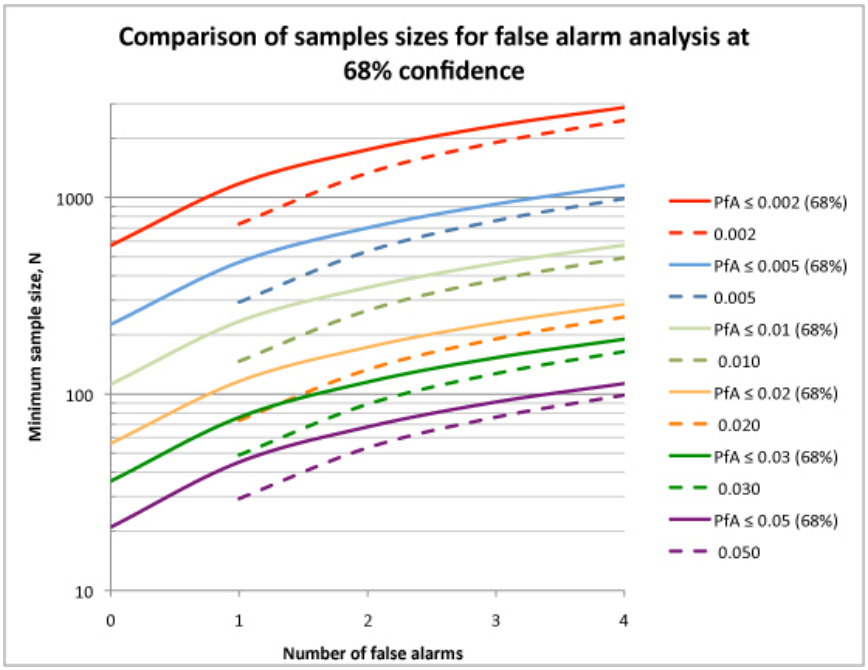


*Figure 4.3-1 Comparison of sample sizes for false alarm analysis at 68% confidence. Solid curves are the exact result from the binomial and dashed lines are the estimates from the Gaussian approximation.*

Examination of the above results shows that the Gaussian approximation slightly underestimates the sample sizes though the estimates improve at lower false alarm thresholds and at larger m. Generally, the Gaussian underestimates the sample size by ~ 15-35% which is remarkably good considering the disparity between the Gaussian and binomial representations are the cases of primary interest.

A similar comparison at 95% confidence is shown below.

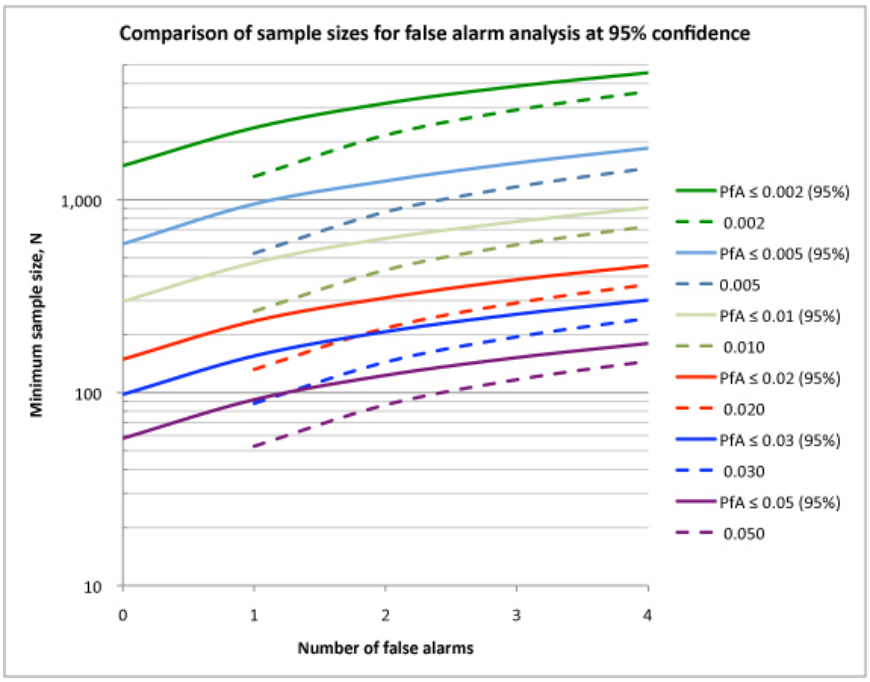


*Figure 4.3-2 Comparison of sample sizes for false alarm analysis at 95% confidence. Solid curves are the exact result from the binomial and dashed lines are the estimates from the Gaussian approximation.*

Examination of the comparative data shows that, at 95% confidence, the Gaussian estimates are systematically a little less than the exact results. However, the results appear to be the same over the whole range the Gaussian approximation underestimating the minimum sample size by ~ 20-40 % over more than two decades span in sample sizes.

# 5 Point estimate

In a "point estimate" presentation the goal is to determine the most probable value of P together with its uncertainty. The latter is represented by the standard deviation of P. The most probable value of P is given, as before for the binomial distribution:

$$\langle P \rangle = \frac{K}{N} \tag{5.1-1}$$

and the variance for the binomial distribution may be found in Wikipedia[11] or in the other statistics references listed earlier:

$$\sigma^2 = NP(1-P) = K\left(1 - \frac{K}{N}\right) \tag{5.1-2}$$

Here K is the number of errors in the data set being analyzed and is usually small compared to N, i.e. K << N, so the approximation used here will be:

$$\sigma \approx \sqrt{K} \tag{5.1-3}$$

and note the similarity to the Poisson and Gaussian distributions. The point estimate is generally utilized only when N is large and the Gaussian is a good approximation in this regime.

In this presentation the goal is to establish the minimum sample size that provides a standard deviation that is small compared to the most probable value. The relative error will be defined by choice of a parameter alpha $\alpha$ below:

$$\alpha \equiv \frac{\sigma}{K} \tag{5.1-4}$$

This parameter, $\alpha$, is set according to the quality or accuracy desired in the result. It is the relative error in the estimation of P. It is a choice made in the process of planning an experiment and serves a role somewhat similar to the confidence in the threshold approach. Choosing $\alpha$=10% provides a reasonably precise estimate of $P_D$ and reducing that to $\alpha$=5% is even more precise, but at the expense of a much more challenging measurement scope. Solving 5.1-4 for N for the detection application* (using 5.1-1 and 2.1-1):

$$N_D = \frac{1}{[1-P_D]\alpha^2} \tag{5.1-5}$$

[11] http://en.wikipedia.org/wiki/Binomial_distribution

* The sample size is determined for the case where $\alpha$ is the relative error in $P_{FN}$, i.e. ratio of standard deviation to the interval $[1-P_D]$. If it is desired to obtain a relative error for $P_D$ itself then $[1-P_D]$ in (5.1-4) should be replaced by $P_D$.

And for the false alarm application:

$$N_{fA} = \frac{1}{P_{fA}\alpha^2} \quad (5.1\text{-}6)$$

Examples of minimum sample sizes for α=20%, 10%, and 5% are shown in the figures below.

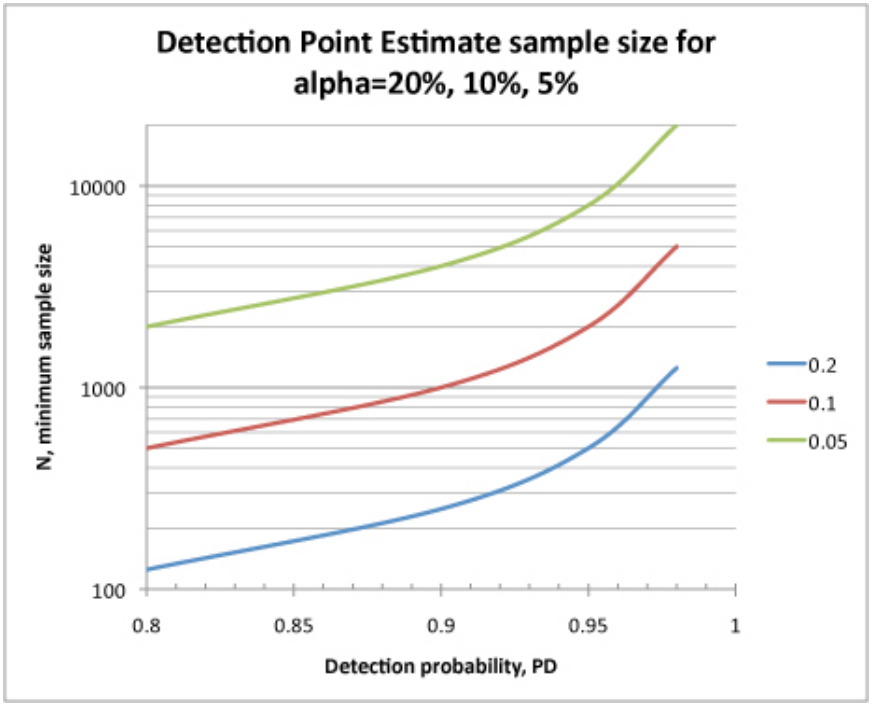


*Figure 5.1-1 Minimum sample size for point estimate vs. detection probability. Curves shown are for α= 20%, 10% and 5%.*

Examination of the figure shows that if a point estimate is desired with only ~ 10% uncertainty then the sample sizes are quite large. For $P_D$=0.8 and 0.9 with 10% error the sample sizes are in the range 500-1000 as compared to the threshold analysis in Fig 3.1-3 where tens of samples were sufficient even at 95% confidence.

A similar comparison can be made for the false alarm analysis and that is shown in the figure below.

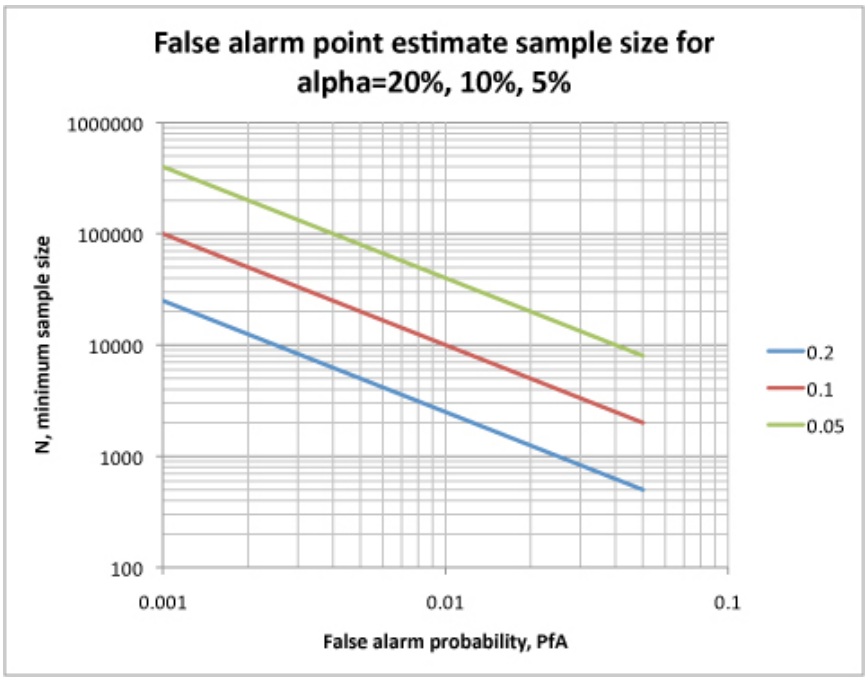


*Figure 5.1-2 Minimum sample size for point estimate of false alarm probability, $P_{fA}$. Curves shown are for $\alpha$= 20%, 10% and 5%.*

Examination of the figure shows the sample size requirements for the point estimate are much larger than were required for the threshold estimates. Comparison with Fig 3.3-3 show that only hundreds of trials or sometimes a few thousand were required even for 95% confidence when the threshold value was considered adequate. For the point estimate over the range of $P_{fA}$ considered the sample size requirements are many tens of thousands.

# 6 Conclusions

Analysis here assumes that a data set is made up of independent random trials. Specifically:

- The sequence of targets of interest is random
- The sequence of interfering materials is random
- The sequence of interfering radiation sources is random
- The time intervals and composition are uncorrelated with any target objects to the extent possible.

The point estimate provides a very precise estimate of system performance but requires very robust statistics, i.e. very large sample size. On the other hand the "threshold" analysis is more ambiguous and establishes only that the performance falls somewhere within an acceptable range with a known probability (confidence). This approach is much less demanding in the statistics required and useful performance measures may be obtained with much smaller sample size. In many applications the results provide

adequate precision and are capable of determining the probability that a given system meets a performance requirement without determining the performance, P, explicitly.

Overall, that the Gaussian approximation provides reasonably accurate estimates of minimum sample size over a range of detection thresholds and limiting false alarm thresholds. Two variations have been presented for application to analysis of detection data (when contraband is present) and to false alarm data (when contraband is not present). These estimates have been compared to the exact binomial formulation over a range of detection thresholds from 0.8 to 0.98 and over a range of false alarm thresholds from 0.05 to 0.002. The Gaussian formulation presented above typically underestimates the minimum sample size by ~ 15-40% but no more than that over the range studied. That conclusion is supported at confidence levels 68% and 95%.